Erling B. Andersen

The Statistical Analysis of Categorical Data

Third Edition

With 41 Figures

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Professor Dr. Erling B. Andersen
Department of Statistics
University of Copenhagen
Studiestræde 6
DK-1455 Copenhagen K
Denmark

ISBN-13 : 978-3-642-78819-2 e-ISBN-13 : 978-3-642-78817-8
DOI: 10.1007/978-3-642-78817-8

42/2202-5 4 3 2 1 0 - Printed on acid-free paper

To Ellen

for love, inspiration and infinite patience

Preface

The aim of this book is to give an up to date account of the most commonly uses statistical models for categorical data. The emphasis is on the connection between theory and applications to real data sets. The book only covers models for categorical data. Various models for mixed continuous and categorical data are thus excluded.

The book is written as a textbook, although many methods and results are quite recent. This should imply, that the book can be used for a graduate course in categorical data analysis. With this aim in mind chapters 3 to 12 are concluded with a set of exercises. In many cases, the data sets are those data sets, which were not included in the examples of the book, although they at one point in time were regarded as potential candidates for an example.

A certain amount of general knowledge of statistical theory is necessary to fully benefit from the book. A summary of the basic statistical concepts deemed necessary prerequisites is given in chapter 2.

The mathematical level is only moderately high, but the account in chapter 3 of basic properties of exponential families and the parametric multinomial distribution is made as mathematical precise as possible without going into mathematical details and leaving out most proofs.

The treatment of statistical methods for categorical data in chapters 4 to 12 is based on development of models and on derivation of parameters estimates, test quantities and diagnostics for model departures. All the introduced methods are illustrated by data sets almost exclusively from Danish sources. If at all possible, the data source is given.

Almost all statistical computations require the use of a personal or main frame computer. A desk calculator will only in few cases suffice. As a general rule the methods in chapters 4 to 7 are covered by standard statistical software packages like SAS, BMDP, SPSS or GENSTAT. This is not the case for the methods in chapters 8 to 12. Søren V.

Andersen and the author have developed a software package for personal computers, called CATANA, which cover all models in chapters 8 to 12. This package is necessary in order to check the calculations in the examples or to work through the exercises. Information on how to obtain a diskette with CATANA, which will be released in early 1990, can be obtained by writing to the author.

A fair share of the examples and exercises are based on the Danish Welfare Study and I wish to thank the director of this study professor Erik J. Hansen, who through the Danish Data Archive put the data file from the Welfare Study to my disposal, and has been extremely helpful with extra information on the data.

Part of the book was written during visits to the United States and France. I wish to thank first of all Leo Goodman, but also Peter Bickel, Terry Speed, Jan de Leeuw, Shelby Haberman, Peter McCullogh, Darrell Bock, Clifford Clogg, Paul Holland, Robert Mislevy and Murray Aitkin in the United States and Yves Escoufier, Henri Caussinus and Paul Falguerolles in France for stimulating discussions. Many other persons have contributed to the book through discussions and criticism. It is impossible to name all, but the help of Svend Kreiner, Nils Kousgaard and Anders Milhøj is appreciated.

I also wish to thank the Danish Social Science Research Council, who financed my visits to the United States and France.

The book would never have been a reality without the care and enthusiasm with which my secretary Mirtha Cereceda typed and retyped the manuscript many times. I owe her my most sincere thanks for a very competent job.

Copenhagen, October 1989

In this second edition a number of errors have been corrected and a new chapter 13 on computer packages been added.

Copenhagen, June 1991

In the third edition all figures have been redrawn and a few minor errors corrected.

Copenhagen, October 1993
Erling B Andersen

Contents

1.	**Categorical Data**	**1**
2.	**Preliminaries**	**9**
2.1	Statistical models	9
2.2	Estimation	10
2.3	Testing statistical hypotheses	14
2.4	Checking the model	19
3.	**Statistical Inference**	**25**
3.1	Log–linear models	25
3.2	The one–dimensional case	29
3.3	The multi–dimensional case	43
3.4	Testing composite hypotheses	55
3.5	The parametric multinomial distribution	62
3.6	Generalized linear models	70
3.7	Solution of likelihood equations	74
3.8	Exercises	82
4.	**Two–way Contingency Tables**	**89**
4.1	Three models	89
4.2	The 2x2 table	94
4.3	The log–linear parameterization	105
4.4	The hypothesis of no interaction	108
4.5	Residual analysis	120
4.6	Exercises	121
5.	**Three–way Contingency Tables**	**131**
5.1	The log–linear parameterization	131
5.2	Hypothesis in a three–way table	135
5.3	Hypothesis testing	143
5.4	Decomposition of the test statistic	164
5.5	Detection of model departures	167
5.6	Exercises	173
6.	**Multi–dimension Contingency Tables**	**179**
6.1	The log–linear model	179
6.2	Interpretation of log–linear models	181
6.3	Search for a model	190
6.4	Diagnostics for model departures	199
6.5	Exercises	202

7.	**Incomplete Tables, Separability and Collapsibility**	212
7.1	Incomplete tables	212
7.2	Two–way tables and quasi–independence	216
7.3	Higher order tables. Separability	220
7.4	Collapsibility	228
7.5	Exercises	234
8.	**The Logit Model**	**239**
8.1	The logit–model with binary explanatory variables	239
8.2	The logit model with polytomous explanatory variables	254
8.3	Exercises	265
9.	**Logistic Regression Analysis**	**269**
9.1	The logistic regression model	269
9.2	Regression diagnostics	286
9.3	Predictions	303
9.4	Polytomous response variables	305
9.5	Exercises	311
10.	**Models for the Interactions**	**320**
10.1	Introduction	320
10.2	Symmetry models	320
10.3	Marginal homogeneity	329
10.4	Models for mobility tables	333
10.5	Association models	336
10.6	RC–association models	345
10.7	Log–linear association models	352
10.8	Exercises	354
11.	**Correspondence Analysis**	**362**
11.1	Correspondence analysis for two–way tables	362
11.2	Correspondence analysis for multi–way tables	387
11.3	Comparison of models	397
11.4	Exercises	402
12.	**Latent Structure Analysis**	**406**
12.1	Latent structure models	406
12.2	Latent class models	407
12.3	Continuous latent structure models	411
12.4	The EM–algorithm	424
12.5	Estimation in the latent class model	426
12.6	Estimation in the continuous latent structure model	437
12.7	Testing the goodness of fit	453

12.8 Diagnostics 462
12.9 Score models with varying discriminating powers 473
12.10 Comparison of latent structure models 476
12.11 Estimation of the latent variable 479
12.12 Exercises 481

13. Computer Programs 495

References 504

Author Index 519

Subject Index 524

Examples with Data 532